Maximum Entropy Based Prediction of UMLS Semantic Types

Kirk Baker*, Laura Cameron, Allison Arinaga, Jade Blevins, Krishna Collie, Agnes Demianska, Rick Ikeda, Erin Lee, Jessica Lobo, Judy Riggie, Jonathan Sears and George Santangelo*

Office of Research Information Systems, *Office of Portfolio Analysis

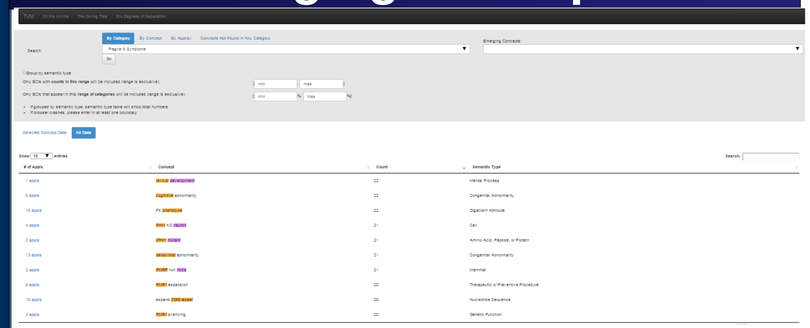
Background

Structured vocabularies such as the RCDC Thesaurus or MeSH are useful for processing and organizing large volumes of textual information. One challenge in using structured vocabularies is keeping them up to date: new terms continually appear in the literature and must be incorporated. Statistical natural language processing techniques can be used to identify candidates for inclusion and classify them semantically. We describe an approach to predict the most likely semantic class of candidate terms as well as practical applications of the approach.

Methods

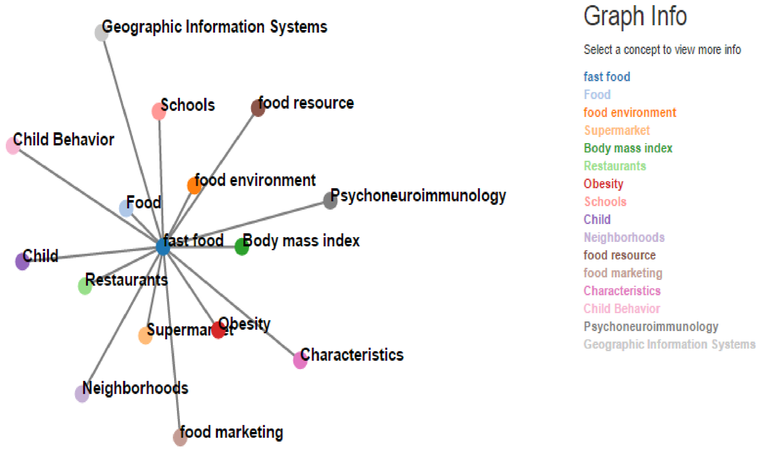
- Utilized a maximum entropy framework (*aka* multinomial logistic regression) customized from the openNLP MAXENT Java implementation
- Training data (220,000 entries) were extracted from MeSH and augmented with selected WordNet files (noun.body, noun.person)
- Each item in the training data was converted into an outcome (response variable) and a set of predictive features (independent or predictor variables)
- Features consisted of a head word (the linguistic head of a phrase, with suffixes removed) and bigrams (overlapping two-word sequences of words)
- Example:
  'soy cheese'
  outcome=Food head=cheese bigram=# soy bigram=soy cheese bigram=cheese #
- 75% accuracy over 133 UMLS semantic groups; baseline 13%

Emerging Concepts



The Emerging Concept Identifier enables searching and displaying candidate terms, relating them to NIH projects, and filtering by semantic type. Highlighting relates terms to the RCDC thesaurus.

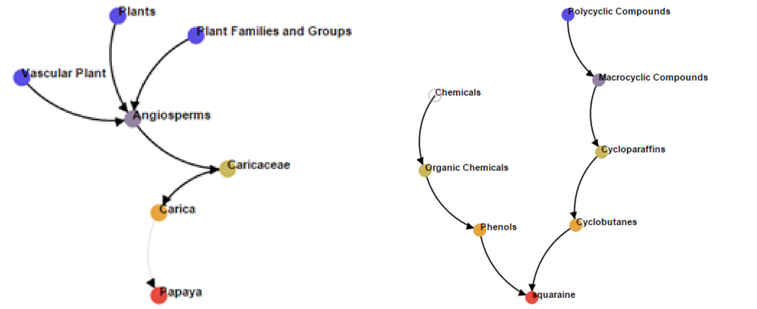
Concept Similarity



Predicted semantic types can be applied to existing RCDC concepts and used to identify or organize contextually similar concepts. Here, concepts that are contextually related to "fast food" in NIH projects are displayed. Contextual similarity was derived from a Chi-Square measure of concept co-occurrence.

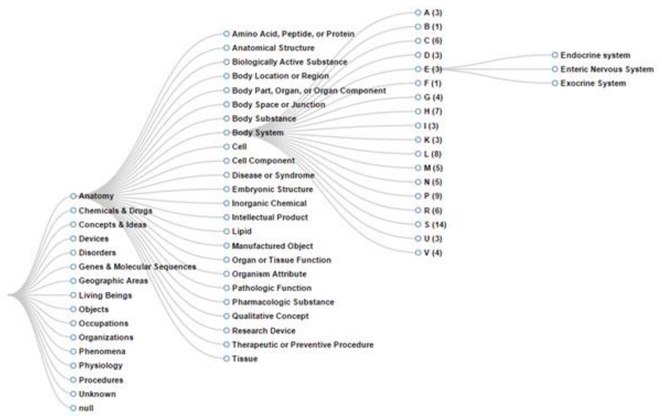
A similar display for concepts related to "Crime" is shown. The length of the edge between nodes corresponds to the strength of the relationship between concepts (shorter edges indicate a stronger contextual relationship).

Inferring Hierarchical Structure



Predicted semantic types can be used to infer an existing or emerging concept's location within a taxonomy. Here, potential hierarchical structures for "Papaya" and "squarine" are shown. These structures were calculated by selecting the most similar semantic type of parent concepts in the RCDC thesaurus.

Structured Thesaurus Browsing



An alternate view of the RCDC Thesaurus, obtained by browsing the predicted semantic type hierarchy of all concepts.

Future Work

We plan to incorporate contextual similarity into the prediction of semantic types and explore the utility of this approach for automated document classification.